A novel strategy for the development of vaccines for SARS-CoV-2 (COVID-19) and other viruses using AI and viral shell disorder

Gerard Kian-Meng Goh,[1,*] A. Keith Dunker,[2] James A. Foster,[3,4] and Vladimir N. Uversky[5,6]

[1] BioComputing, Singapore, Republic of Singapore

[2] Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, Indiana, USA.

[3] Department of Biological Sciences, University of Idaho, Moscow, Idaho, USA.

[4] Institute for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, Idaho, USA

[5] Department of Molecular Medicine, USF Health Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, Tampa, FL, USA.

[6] Laboratory of New Methods in Biology, Institute for Biological Instrumentation of the Russian Academy of Sciences, Federal Research Center "Pushchino Scientific Center for Biological Research of the Russian Academy of Sciences", Pushchino, Moscow region, Russia

[*]Corresponding author

Email addresses:

GKMG: gohsbiocomputing@yahoo.com

AKD: kedunker@iupui.edu

JAF: foster@uidaho.edu

VNU: vuversky@health.usf.edu

Abstract

A model that predicts levels of coronavirus (CoV) respiratory/fecal-oral transmission potentials based on the outer shell hardness has been built using neural network (artificial intelligence, AI) analysis of the percentage of disorder (PID) in the nucleocapsid, N, and membrane, M, proteins of the inner and outer viral shells, respectively. Based mainly on the PID of N, SARS-CoV-2 is categorized as having intermediate levels of both respiratory and fecal oral transmission potential. Related to this, other studies have found strong positive correlations between virulence and inner shell disorder among numerous viruses, including Nipah, Ebola, and Dengue viruses. There is some evidence that this is also true for SARS-CoV-2 and SARS-CoV, which have N PIDs of 48% and 50%, and are characterized by case-fatality rates of 7.1% and 10.9%, respectively. The link between levels of respiratory transmission and virulence lies in viral load of body fluids and organ respectively. A virus can be infectious via respiratory modes only if the viral loads in saliva and mucus exceed certain minima. Likewise, a person may die, if the viral load is too high especially in viral organs. Inner shell proteins of viruses play important roles in the replication of viruses, and structural disorder enhances these roles by providing greater efficiency in protein-protein/DNA/RNA/lipid binding. This paper outlines a novel strategy in attenuating viruses involving comparison of disorder patterns of inner shells of related viruses to identify residues and regions that could be ideal for mutation. The M protein of SARS-CoV-2 has one of the lowest M PID values (6%) in its family, and therefore this virus has one of the hardest outer shells, which makes it resistant to antimicrobial enzymes in body fluid. While this is likely responsible for its contagiousness, the risks of creating an attenuated virus with a more disordered M are discussed.

Key words: intrinsic; disorder; protein; nucleocapsid; Nipah; virulence; viral protein; protein structure; protein function, shell; covid;coronavirus; ebola;vaccine; immune; antibody;shell; nucleocapsid; nucleoprotein; matrix;attenuate;

Introduction

SARS-CoV-2 (COVID-19)

The coronavirus infectious disease 2019 (COVID-19) is caused by the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) [1]. The first sign of the SARS-CoV-2 spread was reported in December 2019 at a Wuhan seafood market in China that also sold live animals [2-5]. Since then, COVID-19 has become an extremely serious pandemic with confirmed global infections and deaths moving quickly towards 3.5 million and 250,000 respectively. The current case-fatality rate of COVID-19 is 7.1%. SARS-CoV-2 is closely related to SARS-CoV that caused an outbreak in 2002-2003 but was more contagious (it showed the case-fatality rate of 10.9%).

Intermediate levels of both fecal-oral and respiratory transmission potentials predicted for SARS-CoV-2 and SARS-CoV

A model was built before the MERS-CoV outbreak in 2012. This model measured the percentage of intrinsic disorder (PID) of proteins from the viral inner and outer shells, the nucleocapsid, N, membrane protein, M, respectively [6,7]. Upon the tabulation of the PIDs, the CoVs were clustered into three groups (**Table 1**). The first group, group A, consists of coronaviruses that have lower fecal-oral but higher respiratory transmission potentials. Group B are CoVs that have intermediate levels of fecal-oral and respiratory potentials. Lastly, group C includes viruses that have higher fecal-oral, but lower respiratory transmission potentials. The clustering is based mainly on PID values of corresponding N proteins, even though later statistical analysis detected that PID values of M proteins also slightly contribute to the categorization. The model indicated that SARS-CoV (PID_M = 8%, PID_N = 50%) falls into category B, which is indicative of intermediate levels of both fecal-oral and respiratory transmission levels.

Chances for further validation of the model came with the 2012 MERS-CoV and the 2019 COVID (SARS-CoV-2) outbreaks. Proteomic and genetic analyses of the viruses indicates that MERS-CoV

(PID_M = 9%, PID_N = 44%) and SARS-CoV-2 (PID_M =6%, PID_N = 48%) belong to the categories C and B respectively, according to the model already established [2-3,7,8]. These results are consistent with clinical and other observations. MERS-CoV originated from a zoonotic transmission event of a camel coronavirus. Since fecal-oral transmission is generally the most efficient mode of transmissions among farm animals, including camels, this could evolutionary account for the high fecal-oral transmission potential of MERS-CoV as predicted by the model [9,10]. SARS-CoV-2, on the other hand, falls into the same category B as its closely related cousin, SARS-CoV, with the intermediate levels of both respiratory and fecal-oral transmission potentials.

Reason for higher SARS-CoV-2 infectivity: Harder outer shell, greater resilience and, thus, greater contagiousness

The model detects, however, something else odd about SARS-CoV-2 and shows that its outer shell is among the hardest outer shells (i.e., low PID_M = 6%) within the CoV family. As the outer shell plays the greatest role in protecting the virion, it is therefore likely that SARS-CoV-2 should be more resistant to the antimicrobial enzymes found on the tongue and skin or in saliva, mucus, and other body fluids. As a result, the body is capable of shedding more viral particles. This might account for the fact that SARS-CoV-2 is more contagious than SARS-CoV.

Inner shell disorder (N protein): A novel vaccine target

Enigmatically, in other viruses analyzed by the model, such as Ebolavirus (EBOV), Nipah virus (NiV), Dengue (DENV) and flaviviruses, the inner shell disorder was also found to be correlated with the level of virulence [7,10-13]. It is for this reason, the inner shells have been seen as potential vaccine targets. Our previous paper has briefly examined the relationship between the virulence and respiratory transmission potentials via analysis of the inner shell disorder and protein binding promiscuity [3]. This paper will examine in greater details the strategy of using

nucleocapsid (N) as a vaccine target not just for SARS-CoV-2, but also for a variety of related and unrelated viruses.

Disorder prediction tools and reproducibility of the model

Predictors of protein intrinsic disorder and other sequence analysis tools

A major concept used in the study is the phenomenon of protein intrinsic disorder, which refers to proteins or protein regions that have no unique 3D structure. While protein structures are linked to protein functions via classic protein structure-function paradigm (where unique protein sequence encodes unique 3D structure that is responsible for unique biological function), protein disorder, likewise, has been linked to myriads of protein functions [14-16]. Tools have been developed to predict disorder, since disordered and ordered proteins are characterized by specific and therefore predictable features of their amino acid sequences. The first of these disorder predictors is the PONDR® VLXT (www.pondr.com), which is a neural network (artificial intelligence: AI) trained to recognize disordered and ordered sequences [17-19].

PONDR® VLXT has been highly successful when used to study viral proteins, especially viral shell proteins, since many of these proteins are structural proteins held together by protein-protein or protein-RNA/DNA interactions, and PONNDR® VLXT is known to be one of the best predictors that can take into account these factors [20,21]. It is for these reasons PONDR® VLXT has been successfully used in the study of a large variety of viruses including human immunodeficiency virus (HIV), herpes simplex virus (HSV), hepatitis C virus (HCV), NiV, EBOV, 1918 HIN1 influenza A virus, CoVs, and flaviviruses including yellow fever (YFV), Zika (ZIKV), DENV [2-3,6-8,10-13,23-31].

An important number that is used as a yardstick to measure the level of disorder in a protein is the percentage of intrinsic disorder (PID), which is defined as the number of residues predicted to be disordered divided by the total number of residues in a protein. The sequences are available at UniProt

(http:www.uniprot.org) and NCBI (https://www.ncbi.nlm.nih.gov/nuccore/MN908947). Relational database is used to store disorder and sequence information [22]. Statistical calculations using multivariate analysis were done using R statistical package [28]. Basic Alignment Search Tools for Proteins (BLASTP) is available at NCBI (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins).

Revisiting the CoV shell disorder model

As aforementioned, the CoV shell disorder model is based on the disorder levels of two major shell proteins: M (membrane, outer shell) and N (nucleocapsid, inner shell) [7,29]. **Table 1** shows the grouping of the three category of coronaviruses. **Figure 1** tells us that the SARS-CoV-2 is odd as its outer shell (M) is among the hardest in the family given its low PID_M value, which is likely indicative of greater resistance of this virus to antimicrobial enzymes found in the saliva, mucus, tongue, and skin [2,3]. This characteristic is believed to cause greater shedding of viral particles that is responsible for its greater contagiousness. There is, however, at least one other "competing" theory that could also accounts for the greater contagiousness of SARS-CoV-2. This involves the discovery that the spike (S) glycoprotein binds 20-30 times more tightly to the angiotensin converting enzyme-2 (ACE2) receptor than in the case of SARS-CoV [30]. While this finding does not discount, in any way, the results found in **Table1** and **Figure 1**, it raises questions pertaining to the true cause of its contagiousness. With a high probability, both of these factors contribute to the higher transmission potential of SARS-CoV-2.

Incoming COVID-19 data supporting the CoV shell model

Incoming clinical data are increasingly providing compelling evidence that the SARS-CoV-2 sheds large quantities of infectious particles. Heavy viral shedding has been detected on the first day when the patients showed the slightest symptom [31] or even no symptom [32]. The heavy shedding

lasted until the days of recovery. This shedding was observed to be much greater among COVID-19 patients than those infected with SARS-CoV [31]. Large amounts of virions are not confined to the respiratory tract and spread in a form of respiratory droplets, but can also be found in fecal matters. The infectious particles were observed to be active for a prolonged period of time [33]. Needless to say, these are consistent with all the predictions that our model has made, including the greater probability that the viral particles will remain active for a longer period of time, since SARS-COV-2 has one of the hardest outer shells in its family, and that the SARS-COV-2 has intermediate levels of both respiratory and fecal-oral transmission potentials.

Lineage, intermediary hosts, and PID_M and PID_N values

A glance at the PID_N values of various CoVs in **Figure 1** tells us that the neighboring viruses with similar PID_N values are bat coronaviruses. Not only this corroborates with phylogenetic studies showing that one of the bat CoVs (RATG13) has as much as 96% similarity to SARS-CoV-2 [4,5], it highlights the evolutionary connection between this virus and bat CoV. One interpretation would be that the PID_N in the range of 47-48% defines the optimal respiratory transmission potential to spread among the bats. If this is the case, then the N protein from SARS-CoV-2 did not evolve much, if any, and is ton too different from its ancestral counterpart in bat CoV, even if there is an intermediary host before the appearance of CoV capable of infecting humans. This is in sharp contrast to its close cousin, SARS-CoV, which has PID_N of 50%, which could imply that the N protein from SARS-CoV probably had time to evolve in an intermediary host, such as civet cat.

A different story can be found while looking at the PID_M values in different viruses. This is not contradictory, as different proteins affect organisms differently. For instance, a virus may need more immediate protection by evolving a harder outer shell (more rigid M protein), if it moves to a new host species that has stronger antimicrobial enzymes in its saliva [2,3]. **Figure 1** shows that very few other CoVs have such low PID_M values as SARS-CoV-2 and, unlike disorder levels in N proteins, none of the ARS-CoV-2 bat cousins are close to such low PID_M. The closest counterparts

found in this sample include canine-respiratory CoV and bovine CoV. This is likely a reflection of the possibility that SARS-CoV-2 M protein had evolved in an intermediary host. Indeed, there have been much debate with regard the possibility of a snake or pangolin could serve as intermediary host for SARS-CoV-2 [34,35]. Furthermore, SARS-CoV-2 is now know to infect domesticated animals, such as cats and dogs [36]. Our model seems to imply that there is probably a non-bat intermediary host that the virus had for just enough time for its M protein to evolve, but its N protein probably needed a longer time to evolve outside the original bat host, and our data show that the N protein did not have a chance to evolve in an animal host other than bats.

Links between virulence and inner shell disorder

Links among modes of transmission and virulence with inner shell (N) disorder

Correlations between inner shell disorder and virulence have been previously discovered among a large variety of viruses, such as NiV, EBOV and flaviviruses, such as Zika (ZIKV), Dengue (DENV), and yellow fever (YFV) viruses. **Figure 2** represents the links between case-fatality rates (CFR) and inner shell disorder levels among a variety of viruses. While **Figure 2A** show some evidence of a link between SARS-CoV/SARS-CoV-2 virulence and PID_N, **Figures 2B** [12], **2C** [13], **2D** [10] and **2E** [11] show correlations between inner shell disorder and CFRs of NiV, filoviruses (including EBOV), DENV, and flaviviruses respectively. While noting that DENV (**Figure 2D**) is a flavivirus (**Figure 2E**), **Figures 2A**, **2B**, **2C** and **2D-E** represents groups of RNA viruses that are not related between groups. For this reason, somewhat different nomenclatures and abbreviations are used for the major inner shell protein chosen from different family of viruses. The nucleocapsid, nucleoprotein, and capsid of NiV, EBOV, and DENV are denoted by N, NPm and C respectively [29]. As also seen in **Figure 1**, the coefficients of determination (r^2) of greater than 0.25 for the various viruses reveal strong correlations between virulence and inner shell disorder.

PID_N and CFR values of SARS-CoV and SARS-CoV-2 suggest a link between virulence and N disorder

Figure 2A shows that there is a link between SARS CFR and PID_N. Reasonably and commonly measured CFR for SARS-CoV-2 is in the range of 3-7%. SARS-CoV-2 is generally believed to be less severe than SARS-CoV, which has a CFR of about 10% [37-39]. However, it should be noted that establishing a link between virulence and PID_N among CoV is an intricate matter, since CoVs are genetically diverse and have 4 genera. Furthermore, CoVs infects a large variety of animal hosts with different virulence and often use different host receptors. For instance, both SARS-CoV and SARS-CoV-2 enter the host cell using ACE2 receptor, while MERS-CoV uses dipeptidyl peptidase-4 DPP4 [7,30]. Even more confusingly, MERS-CoV has a 35% CFR for humans, but is generally harmless to its predominant camel host [9]. Incidentally, SARS-CoV and SARS-CoV-2 lies within the same lineage of β-CoVs and share close to 80% in genomic sequence identity, whereas MERS-CoV has only 50% sequence similarity to SARS-CoV-2 [40]. For these reasons, MERS-CoV cannot be included in the correlation.

HCOVs that are distantly related to SARS-CoV and SARS-CoV-2 should also not be included. For instance, HCOV-NL63 should not be included even though it also enters human host using the ACE2, but in a different manner from SARS-CoV/SARS-CoV-2 [41,42]. HCOV-NL3 and HCOV-229E are α-CoVs [43], while SARS-CoV, SARS-CoV-2 and MERS-CoV are in the β-CoV genus [29,35]. Because of the distant relationship between SARS-CoV/SARS-CoV-2 and HCOV-229e/NL63, their sequence similarity is much lower than 50%, similar to what we have seen in the case of MERS-CoV and SARS-CoV-2. In fact, a sequence similarity study of the S proteins of SARS-CoV-2 and HCOV-NL63 has shown the percentage of identity to be as low as 14% [40]. Given these, it is only reasonable that HCOV-229e/NL63 should not be also included into this analysis.

Link between virulence and modes of transmission: Promiscuous protein-protein binding via inner shell disorder

Links between modes of transmission and N disorder: NiV and CoV

Upon the examination of **Table 1** and **Figure 2E**, one can see that there is also a link between the transmission modes and virulence and the inner shell disorder. This link has also been found in NiV. While **Figure 2B** shows correlation between NiV PID_N and virulence, the NiV 1998-9 strain did not involve human to human transmission, unlike the other strains. Transmission during the Malaysian 1998-1999 outbreak involved mainly farmers, who were infected during handling of pigs and their fecal matters, whereas the human-to-human spread that involve respiratory transmission did commonly occur in other strains [12]. Because the 1998-1999 strain has the lowest CFR and PID_N with the lowest observed respiratory transmission, a positive correlation between PID_N and level of viral respiratory transmission potentials has been detected. Likewise, a positive correlation between N disorder and respiratory transmission potential has already been seen in **Table 1**.

Links among virulence, respiratory transmission potential, and N disorder: Viral loads in the body and in body fluids

The afirementioned link among virulence, modes of transmission, and disorder of N protein may seem like a perplexing enigma, but it is not. When we built the CoV shell model before 2012, the viruses were clustered into levels of respiratory potentials in line with the ordering of PID_N. The reason has to do with the fact that the viral loads found in saliva and mucus of the infected person have to be above certain minimal levels before the virus can becomes infectious via the respiratory modes of transmission [2,3]. Similarly, in the case of virulence, death often occurs when the viral load in the human body crosses a certain threshold [7,11-13,18,24-27] . Inner shell disorder holds the connection for these two phenomena, as the inner shell is intimately involved in the viral replication process.

The "Trojan horse" immune evasion: Replicating quickly via N disorder before the immune system detects its presence

Similar proteins often share similar functions even in unrelated viruses [29]. This is especially the case for the nucleocapsid proteins, which are known to play important and similar roles in the replication of viral particles in the host cell [7,25,44-46]. During these processes, they are known to bind to proteins that are part of the host replicating machinery, and by being more disordered, they show greater binding efficiency, as the disorder is known to be associated with more promiscuous protein-protein interactions [14-16,47]. The ability to replicate quickly before the onset of immune response is part of a "Trojan horse" immune evasion strategy [7]. Often, though, this strategy backfires on the virus, as the high viral load overwhelms the host organs and thus leads to death of the host. This is the reason that correlations of virulence and inner shell disorder are easily found in a large variety of related and unrelated viruses. The ability to rapidly replicate utilizing the advantages of the inner shell disorder is also manifested as greater respiratory potentials, as this will also allow for greater viral load in body fluids, such as saliva and mucus, not just in the blood and internal organs.

Functional similarity of inner shells across viruses and the advantages of more promiscuous binding arising from greater levels of disorder

It has long been known that greater disorder levels in the proteins are associated with the greater ability of proteins to be promiscuous binders; i.e., to interact with a greater variety of partners, including lipids, RNA/DNA, and proteins with higher efficiency [14-16,47]. This trait is likely to be of advantage, when a viral protein needs to bind to host proteins with greater effectiveness, as a part of the process of the hijacking of the host cell machinery. The inner shell proteins play crucial roles in the replication of infectious viral particles and their disorder helps with the rapid replication.

There are plenty of opportunities for protein intrinsic disorder to play vital roles, when binding

between the viral and host proteins are considered. For instance, CoVs N protein transport viral proteins and RNA to areas near the endoplasmic reticulum (ER) and Golgi apparatus, where it helps in the packaging of viral particles [3,25,46]. Likewise, the C protein precursor from a flavivirus would egress and then bind to the membrane of the ER, where it interacts with other viral proteins and RNA for assembly and budding [7,10,11,29]. As for EBOV, its NP is responsible for the building of tube-like structures that facilitates the transportation and budding of viral particles [13, 44]. Lastly, the greater disorder in the NiV N protein provides important means for the more efficient binding of this protein to P and L proteins to form a complex, which becomes an RNA polymerase responsible for the viral RNA replication [12,45]. We can now see that the inner shell proteins of the various viruses often play similar roles that provide for ample opportunities for protein intrinsic disorder to contribute to more efficient interactions with the host and viral proteins, DNA/RNA and lipids. Needless say, the greater efficiency of these interactions leads to quicker replication of viral particles.

Strategies for attenuated vaccine development via analysis of inner shell disorder

Attenuating the virus such that it does not replicate rapidly by increasing order in N

We have seen that inner shell disorder is likely to be responsible for fast replication of many viruses. Since this virulence arises form the increased viral load caused by greater disorder in the inner shell protein, a strategy for developing an attenuated vaccine would be to find ways of reducing disorder of N protein in the case of SARS-CoV-2. It would be most ideal if an attenuated virus could still replicate but only very slowly, such that it will maintain the lowest possible viral load.

Figure 3 shows the RNA binding domain of the N proteins of SARS-CoV-2 and murine hepatitis virus (MHV) with the disordered regions represented in red. It can be observed that the N protein

from SARS-CoV-2 (**Figure 3A**) contains noticeably more disordered regions than the N protein from MHV (**Figure 3B**), even though the MHV PID_N (46%) is just moderately lower than the SARS-CoV-2 PID_N (48%). Since **Figure 3** represents the structures of the RNA binding domain, it is likely that the greater disorder differences lie in this region. The reason that MHV is chosen in **Figure 3B** is that its data is easily accessible in PDB. There are, of course, CoVs with lower PID_N values as seen in Table 1. One illustrative example is shown in **Figure 4** representing data for the close cousin of MHV, HCOV-HKU1, which has the lowest PID_N among the CoV samples shown in **Table 1**.

Choosing regions and residues to mutate using comparative disorder pattern

We have seen that N disorder is likely a large factor in virulence, as it might contributes to greater viral load in the body. The next question is then: How do we mutate SARS-CoV-2 such that it becomes optimally attenuated without it dying out from being totally unable to replicate. A strategy would be to compare the N sequences and disorder patterns of a CoV with a lower PID_N to those of SARS-CoV-2. **Figure 4A** represents the PONDR® VLXT plots for SARS-CoV-2 (blue), SARS-CoV (red), and HCOV-HKU1 (dashed green). Disordered regions are denoted by the PONDR® VLXT scores of 0.5 or above. Keeping in mind that PID_N of SARS-CoV-2, SARS-CoV, and HCOV-HKU1 are 48%, 50%, and 37% respectively, we can see plenty of regions, which are disordered in SARS-CoV-2 but are not in HCOV-HKU1. These regions make good potential targets in the development of vaccine candidates. It should be reminded that HCOV-HKU1 was chosen because of its lowest PID_N, but there could also be other good N protein with low PID that can be used to compare with SARS-CoV-2 N disorder pattern.

Sequence analysis

Upon a study of **Figure 4A**, one region that sticks out is a disorder peak around locations 74-99, which

is labelled as 'X' in the graph. Apparently, this is a region, where the largest differences in disorder propensity lies, not just between SARS-CoV-2 and HCOV-HKU1, but also between SARS-CoV-2 and SARS-CoV. This area falls within the RNA binding domain approximately at locations 1-140. **Figure 4B** shows a BLASTP (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins) alignment of SARS-CoV and SARS-CoV-2 of part of the N proteins. A large gap in disorder differences can be found around locations 17-27 (**Figure 4A-B**) that is likely to be responsible for the slight difference (48% Vs, 50%) in N PIDs between SARS-CoV-2 and SARS-CoV. The source of the differences can be found in a specific mutation, within the same region. This involves a single amino-acid mutation of S (Serine) to T (Threonine) (**Figure 4B**) that can be seen when we compare SARS-CoV-2 to SARS-CoV. Serine (S) is more polar than threonine(T) and is therefore more disorder inducing. [17-19].

We can also see, in **Figure 4C**, a large number of amino-acid replacements from polar residues to non-polar residues when we compare the sequences of SARS-CoV-2 and HCOV-HKU1 N proteins respectively. Disorder and order inducting residues are generally polar and non-polar respectively [17-19]. All these provide for a wide range of potential mutations available for use in order to induce greater order to SARS-CoV-2 N protein while attempting to attenuate the virus.

The nature of M protein: Hidden risks and feasibility of vaccine development

The risks of genetically manipulating M

We have seen how a more rigid M causes greater contagiousness by allowing the virus to be more resistant to antimicrobial enzymes found in body fluid and also to be more resilient outside the body. It is therefore a possible temptation to contemplate manipulating M so that the attenuated virus to decrease the chances of the vaccine virus or mutated form in spreading. Before we ponder on this possibility, we need to have a better understanding of the risks of increasing M disorder based on what is known about the pathogenesis of viruses with disordered outer shell.

Increasing M PID may place greater risks of morbidity, such as fetal morbidity and reduced effectiveness of vaccine

Figure 5 summarizes the dangers and implications of viruses with higher outer shell. While we have seen that greater inner shell disorder causes virulence, viruses with higher levels of outer shell disorder, on the other hand, can cause higher morbidity. For example, ZIKV can inflict higher morbidity on fetus by causing microcephaly; i.e. small head and brain. **Figure 5A** shows that ZIKV has a higher outer shell disorder (PID_M = 29%) [11], which is exceptionally high especially among its flavivirus cousins. This is in sharp contrast to the DENV PID_M of 10% and although DENV2 causes microcephaly but it does it at a much lower rate [11]. As for the SARS-CoV-2 and SARS-CoV, while more data are needed, it is currently observed that the viruses do generally harm the fetus of a pregnant woman [48]. The main reason for the link between outer shell disorder and morbidity has to do with the greater ability of viruses with higher outer shell disorder to penetrate organs as the placenta and brain [7,10-11,24,26,27]. This is likely the result of greater binding promiscuity resulting from higher disorder.

Yet another risk of increasing levels of disorder in M protein in the attempt to attenuate a virus is that it could backfire and actually thwart the effectiveness of vaccine by reducing the ability of the vaccine to elicit immune response. It has been known that some viruses, including HIV, HSV, and HCV evade the host immune system by having highly disordered outer shells, and thereby decreasing the ability of antibodies to bind firmly to the viral surface glycoproteins [7,24,27]. Further discussion on this immune-evasion mechanism is presented in the next paragraph.

Feasibility of SARS-CoV-2 vaccine development: Encouraging news

A nightmare scenario that many scientists fear is the possibility the SARS-CoV-2 vaccine may never be developed, as highlighted by HIV, HCV and HSV, where the vaccine search has taken about 40, 30 and 100 years respectively with no success. The research in this paper is actually a

spin-off from a parent research that began about 15 years ago [7,18,20,22,23]. The latter has found that the reason that it is difficult to find effective vaccines for these viruses has to do with the way that they are transmitted, and their relationships with sexual transmission that leads to the highly disordered outer shells. As a result of the motions of the outer shell proteins, neutralizing antibodies are not able to bind tightly to the surface proteins. **Figure 5B** provides encouraging news as the outer shell PIDs of HSV, HCV, and HIV look nothing like those of SARS-CoV, SARS-CoV-2 (**Figure 5A**), and all CoV samples that we have examined in **Table 1**. The HIV-1, HCV, and HSV maximal outer shell PIDs reach 70%, 53% and 63% [27] respectively, in contrast to SARS-CoV and SARS-CoV-2 PID_M values of 8% and 6%.

Author Contributions

GKMG conceived the idea, collected, and analyzed the data, and wrote the first draft. VNU helped with the collection and analysis of literature data, reviewed and revised the draft. AKD and JAF reviewed the manuscript and provided the resources necessary for the research.

Conflicts of Interest

GKMG is an independent researcher and the owner of Goh's BioComputing, Singapore. GKMG has also written a book ("Viral Shapeshifters: Strange Behaviors of HIV and Other Viruses") on a related subject. The authors have no other potential conflict of interests.

References

1. WHO, Novel coronavirus (2019-nCoV) https://www.who.int/emergencies/diseases/novel-coronavirus-2019.

2 Goh GK, Dunker AK, Foster JA,Uversky VN. Rigidity of outer shell predicted by protein disorder model sheds light on COVID-19(Wuhan-2019-nCoV) infectivity. Biomolecules. 2020;199: e221.

3. Goh GK, Dunker AK, Foster JA, Uversky VN. Shell disorder analysis predicts greater resilience of the SARS-CoV-2 in body fluids a outside the body. Microb Pathog. 2020;144:104177.

4. Li X, Zai J, Zhao Q, Nie Q, Li Y, Foley BT, Chaillon A. Evolutionary history, potential intermediate animal host, and cross-species analyses of SARS-CoV-2. J Med Virol.

2020:27. doi: 10.1002/jmv.25731.

5. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579:270-273. doi: 10.1038/s41586-020-2012-7.

6. Goh GK, Dunker AK, Uversky VN. Understanding viral transmission behavior via protein intrinsic disorder prediction: Coronaviruses. J Pathog 2012;2012:738590.

7. Goh GK. Viral shapeshifters: Strange behavoirs of hiv and other viruses. Singapore: Simplicity Research Institute: 2017.

8. Goh GK, Dunker AK, Uversky VN. Prediction of intrinsic disorder in MERS-CoV/HCoV-EMC supports a high oral-fecal transmission. PLoS Curr. 2013;5..

9. WHO, Middle Eastern respiratory syndrome coronavirus (MERS-CoV) (https://www.who.int/ news-room/fact-sheets/detail/middle-east-respiratory-syndrome-coronavirus-(mers-cov)

9. Ferguson M, Van Kerkhov MD, Identification of MERS-CoV in dromedary camels. Lancet Infect Dis. 2014;14:93-4.

10. Goh GK, Dunker AK, Uversky VN. Correlating flavivirus virulence and levels of intrinsic disorder in shell proteins: Protective roles vs. Immune evasion. Mol Biosyst. 2016;12:1881-1891.

11. Goh GK, Dunker AK, Foster JA,Uversky VN. Zika and flavivirus disorder: Virulence and fetal morbidity. Biomolecules. 2019b;9: e710.

12. Goh GK, Dunker AK, Foster JA, Uversky VN. Nipah shell disorder, mode of transmission and virulence. Microb Pathog. 2020;141:103976.

13 Goh GK, Dunker AK, Uversky VN. Detection of links between Ebola nucleocapsid and virulence using disorder analysis. Mol Biosyst. 2015b;11: 2337-2344.

14. Tompa P. Intrinsically unstructured proteins. Trends Biochem Sci. 2002;27:527-533.

15. Uversky VN, Gillespie JR, Fink AL. Why are "natively unfolded" proteins unstructured under the physiological conditions? Proteins Struct. Funct. Genet.. 2000;41:415-427.

16.. Wright PE, Dyson HJ. Intrinsically unstructured proteins: Re-assessing the protein structure-paradigm. J Mol Biol. 1999;293:321-331.

17. Li X, Romero, P, Rani M, Dunker AK, Obradovic, Z. Predicting protein disorder for N-, C-, and internal regions. Genome Inform Ser Workshop Genome Inform. 1999;10:30-40.

18. Garner E, Romero P, Dunker AK, Brown C, Obradovic, Z. Predicting binding regions within disordered proteins. Genome Informatics 1999;10,:41-50.

19. Romero P, Obradovic Z, Li,\ X, Garner EC, Brown CJ, Dunker AK. Sequence complexity of disordered protein. Proteins. 2001;42:38-48.

20. Cheng Y, Oldfield CJ, Meng J, Romero P, Uversky VN, Dunker AK, Mining alpha-helix-forming molecular recognition features with cross species sequence alignments. Biochemistry. 2007;46;13468-13477.

21. Oldfield CJ, Cheng Y, Cortese MS, Romero P, Uversky VN, Dunker AK, Coupled folding and binding with alpha-helix-forming molecular recognition elements. Biochemistry. 2005;44: 12454-12470.

22. Goh GK, Dunker AK, Uversky VN. Protein intrinsic disorder toolbox for comparative analysis of viral proteins. BMC Genomics.. 2008;9 Suppl. 2:S4.

23. Xue B, Williams RW, Oldfield CJ, Goh GK, Dunker AK, Uversky VN. Viral disorder or disordered viruses: Do viral proteins possess unique features? Protein Pept Lett. 2010;17:932-951

24. Goh GK, Dunker AK, Uversky VN. A comparative analysis of viral matrix proteins using disorder predictors. Virol J. 2008;5:126.

25. Goh GK, Dunker AK, Uversky VN. Protein intrinsic disorder and influenza virulence: The 1918 H1N1 and H5N1 viruses. Virol J. 2009;6:69.

26. Goh GK, Dunker AK, Uversky VN. Shell disorder, immune evasion and transmission behaviors among human and animal retroviruses. Mol Biosyst. 2015a;11: 2312-23.

27. Goh GK, Dunker AK, Foster JA, Uversky VN. HIV vaccine mystery and viral shell disorder. Biomolecules. 2019;9:178.

28. R Core Team, R: A language and environment for statistical computing. Vienna, Austria. 2019.

29. Acheson, N. *Fundamental of molecular virology*. Wiley: 2007.

30. Shang J, Ye G, Shi K, Hu B, Wang Y, et al. Structural basis of receptor recovnition by SARS-CoV-2, Nature. 2020. doi: 10.1038/s41586-020-2179-y.

31. Wolfel R, Corman VM, Guggemos W, et al. Wirological asscessmant of hospitalized patients with COVID-19. J Mol Bio. 2019;431:1650-1670.

32. Day M. Covid-19: Four fifths of cases are aysmptomatic, China figures indicate. BMJ. 2020:27. doi: 10.1136/bmj.m1375.

33. Wu Y, Guo C, Tang L, Hong Z Q, et al. Prolong presence of SARS-CoV-2 viral RNA in faecal smaples. Lancet Gasctroenterol Hepatol. 2020;5:434-435.

34. Fan HH, Wang LQ, Liu WL, An XP, Liu ZD, He XQ, Song LH, Tong YG. Repurposing of clinically approved drugs for treatment of coronavirus disease 2019 in a 2019-novel coronavirus (2019-nCoV) related coronavirus model. Chin Med J (Engl). 2020; 6. doi: 10.1097/CM9.0000000000000797.

35. Ji W, Wang W, Zhao X, Zai J, Li X. Cross-species transmission of the newly identified coronavirus 2019-nCoV.J Med Virol. 2020;92:433-440. doi: 10.1002/jmv.25682.

36. Shi S, Wen Z, Zhong G, Yang H, et al. Suseptability of ferrets, cats, dogs and other domesticated animals to SARS-Coronavirus 2. Science. 2020. doi:10.1126/science.abb7015.

37. Pompetchara E, Ketloy C, Palaga T. Immune responses in COVID-19 and potential vaccines: A lesson learned from SARS and MERS epidemics. Asian Pac J Allergy Immunol. 2020;38:1-9. doi: 10.12932/AP-200220-0772..

38. Verity R, Okell LC, Dorigatti, et al. Estimates of severity of coronavirus disease 2019: A model-based analysis. Lancet Infect Dis, 2020;S1473-3099:30243-727. doi: 10.1016/S1473-3099(20)30243-7.

39. Rajgor DD, Lee MH, Archuleta S, Bagdascarian N, Quek SC. The many estimates of COVID-19 case-fatality rate, Lancet Infect Dis. 2020;pii:S1473-3099-9. doi: 10.1016/S1473-3099(20)30244-9.

40. Mousavizadeh L, Ghasemi S. Genotype and phenotype of COVID-19: Their roles in pathogenesis. J Microbiol Immunol Infect. 2020. doi: 10.1016/j.jmii.2020.03.022.

41. Huang I, Bosch BJ, Li F, et al. SARS Coronavirus but not Human human coronavirus NL63 utilizes Cathepsin L ot infect ACE2-expressing cells . J Biol Chemt. 2005;281:3198-203

42. Hoffman H, Pyrc K, van der Hoek L, Geir M, Berkhour, Pohlmann S. Human coronavirus NL63 employs the severe acute respiratory syndrome coronavirus receptor fo cellular entry. PNAS. 2005;102:7988-7993. doi: 10.1073/pnas.0409465102.

43. Corman VM, Muth D, Niermeyer D, Drosten C,. Host and sources of endemic coronaviruses. Adv Virus Res. 2018;100:163-188. doi: 10.1016/bs.aivir.2018.01.001.

44. Wantanabe S, Noda T, Kawaoka Y. Functional mapping of the nucleoprotein of the ebola virus. J Vir. 2006;80:3743-51. [

45. Habchi, J.; Longhi, S. Structural disorder within paramyxovirus nucleoproteins and phosphoproteins. *Mol Biosyst* **2012**, *8*, 69-81.

46. McBride R, van Zyl M, Fielding BC The coronavirus nucleocapsid is a multifunctional protein. Viruses. 2014;6:2991-3018. doi: 10.3390/v6082991.

47. Macosay-Castillo M, Marvelli G, Guharoy M, et al. The balancing act of intrinsically disordeeered proteins enabling functions while minimizing promiscuity. J Mol Bio. 2019;431:1650-1670.

48. Luo Y, Yin K. Mnagement of pregnant women infected with COVID-19. Lancet Infect Dis. 2020;S1473-3099:3091-2. doi: 10.1016/S1473-3099(20)30191-2.

Figure legends

Figure 1. PID_M values of SARS-CoV and other CoV. SARS-CoV-2 has among the hardest outer shell as seen by its low PID_M (^%).

Figure 2. Relationships between the case-fatality rate (CFR) and the inner shell disorder. **A**. Link between SARS-CoV/SAR-CoV-2 PID_N and CFR. **B**. Correlation between NiV PID_N and CFR ($p< 0.001$, $r^2 = 0.83$), NiV is of *Paramyxoviridae* family. **C**. Correlation between Filovirus (Marburg Virus and EBOV) NP (nucleocapsid) PID and CFR ($p < 001$, $r^2 = 0.6$). **D**. Correlation between flavivirus DENV C (capsid) PID and CFR ($p < 001$, $r^2 = 0.88$). The abbreviations used are: REBOV (Reston EBOV), BEBOV (Bundibugyo EBOV), SEBOV (Sudan EBOV), ZEBOB (Zaire EBOV). TBEV (Tick-borne encephalitis virus), TBEV-Si (TBEV-Siberia), TBEV-FE (TBEV-Far East), TBEV-Eu (TBEV-Europe) and WNV (West Nile Virus)

Figure 3. 3D crystal structure representation of the RNA binding regions of the CoV N proteins with disordered regions represented in red. **A**. SARS-CoV-2 N protein (6m3m.pdb). **B**. Murine hepatitis virus (MHV) N protein (4hdv.pdb).

Figure 4. Comparative PONDR® VLXT plots of SARS-CoV, SARS-CoV-2 and HCOV-HKU1 N proteins with BLASTP alignments. A. PONDR® VLXT Plots. Regions with VLXT scors of 0.5 and above are disordered. X is a peak of interest. B. BLASTP alignment of SARS-CoV-2 and HCOV-HKU1 N proteins with disorder annotation. C. BLASTP alignment of SARS-CoV-2 and HCOV-HKU1 N proteins with disorder annotation. The RNA-binding region lies around locations 1-145. Much of the differences in disorder can be found in this region of interest.

Figure 5. Implications and risks of viruses with greater disorder at the outer shell protein **A**. Links between fetal morbidity and inner shell disorder. Previous research has found a strong correlation

between fetal morbidity and M disorder among various strains of ZIKV and DENV2 ($p < 001$, $r^2 = 0.8$). B. Comparative PID bar chart of viruses with no vaccines, HIV, HCV and HSV. A high level of disorder is missing in SARS-CoV-2 M protein (see also Figure 5A) , unlike the HIV, HCV and HSV outer shells.

Table 1. Grouping of coronaviruses by mainly N disorder to predict respiratory and fecal-oral transmission potentials (Statistical Analyses:Two-Way ANOVA, $p < 0.00$), ($p< 0.001$, $r^2 = 0.8$) .

Shell Disorder Category	Coronavirus	M PID % (UniProt /Genbank Accession Code)[a]	N PID % (UniProt/Genbank Accession Code)[b]	M PID	N PID	Remarks
A	HCoV-229E	P15422	P15130	23	56	Higher Levels of Respiratory Transmission Lower Levels of Fecal-oral Transmission
	IBV(Avian)	P69606	Q8JMI6	10	56	
B	Bovine	P69704	Q8V432	7.4	53	Intermediate Levels of Respiratory and Fecal-oral Transmission
	PEDV(Porcine)	P59771	Q07499	8	51	
	Canine(Resp.)	A3EXD6	A3E2F7	6.5	51	
	HCoV-OC43	Q4VID2	P33469	7	51	
	SARS-CoV	P59596	P59595	8	50	
	HCoV-NL63	Q6Q1R9	Q6Q1R8	11	49	
	Bat-HKU4	A3EXA0	A3EXA1	16	48	
	SARS-Cov-2	**QHD43419.1[d]**	**QHD43423.2[d]**	**6**	**48**	
	Bats	A3EXD6	Q3LZX4	11.5	47	
	Bat-HKU5	A3EXD6	A3EXD7[b]	11	47	
C	MHV(Murine)	Q9JEB4	Po3416	8	46	Lower Levels of Respiratory Transmission Higher Levels of Fecal-oral Transmission
	MERS-CoV	9(K0BU37	K0BVN3	9	44	
	TGEV(Porcine)	P09175	P04134	14	43	
	Canine(Ent.)	B8RIR2	Q04700	8	40	
	HCoV-HKU1	Q14EA7	Q0ZME3	4	37	

[a]M PID refers to the percentage of intrinsic disorder (PID) of the membrane protein (M). PID is measured by the number of residues predicted to be disordered divided by the total number of disorder. M PID predicts the hardness of the virion. M is considered to be one of the outer shell.
[b]N protein refers to the nucleocapsid protein, which is an inner shell protein,
[c]SARS-CoV-2
[d]Genbank accession code, https://www.ncbi.nlm.nih.gov/nuccore/MN908947

Figure 1

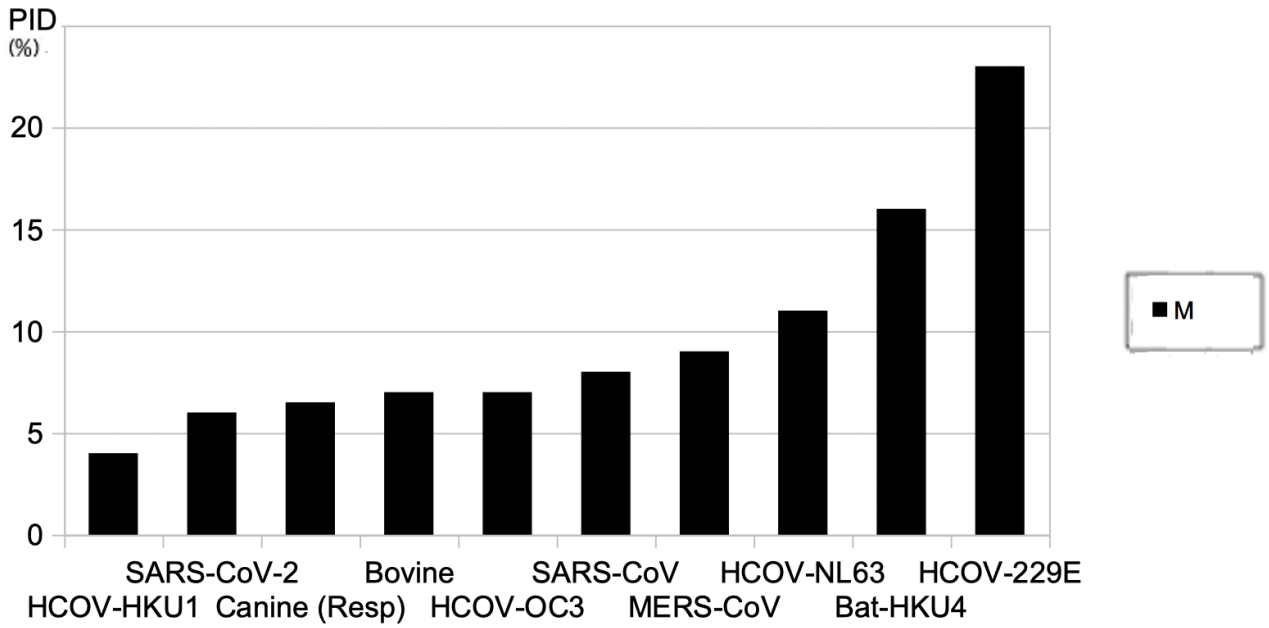
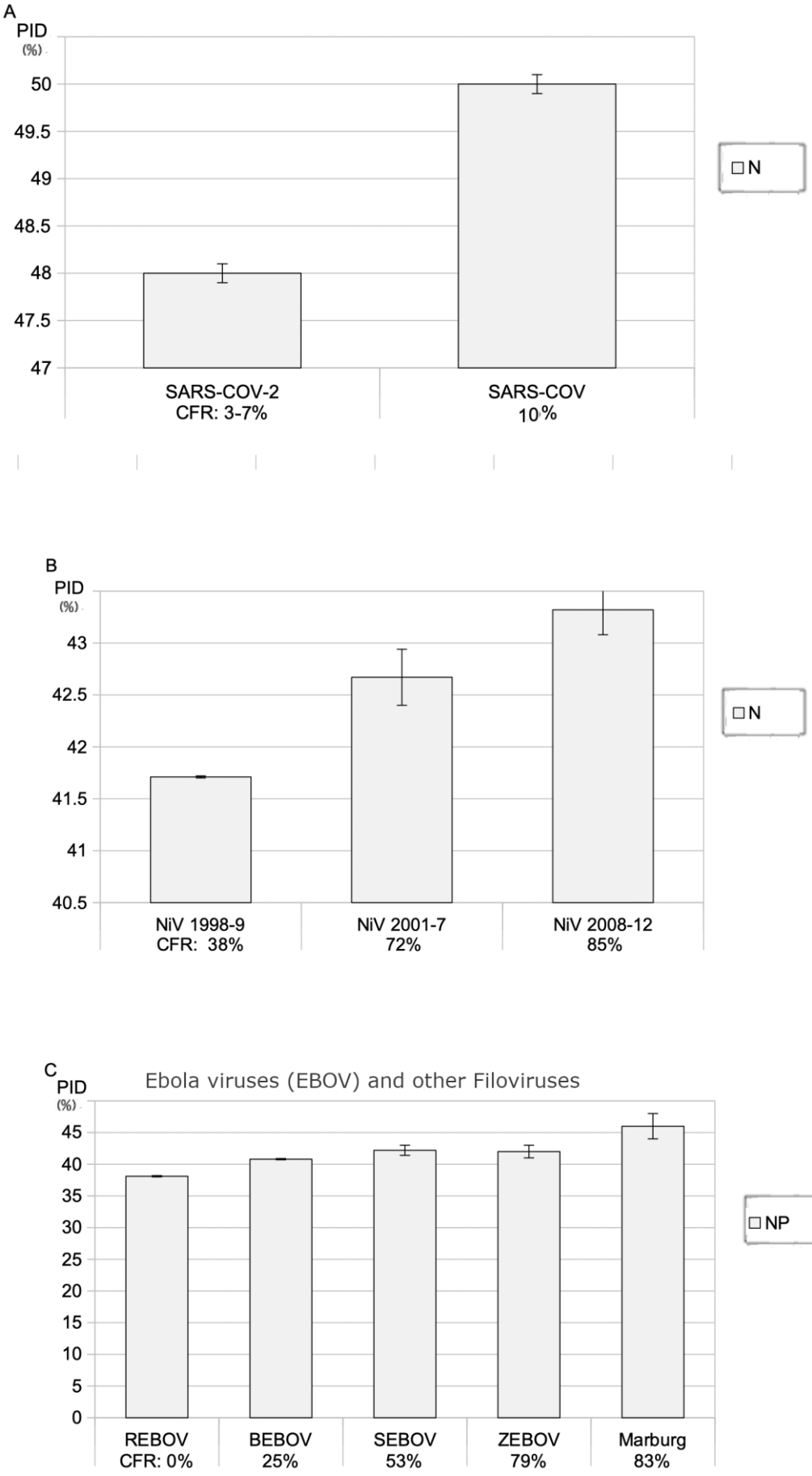


Figure 2



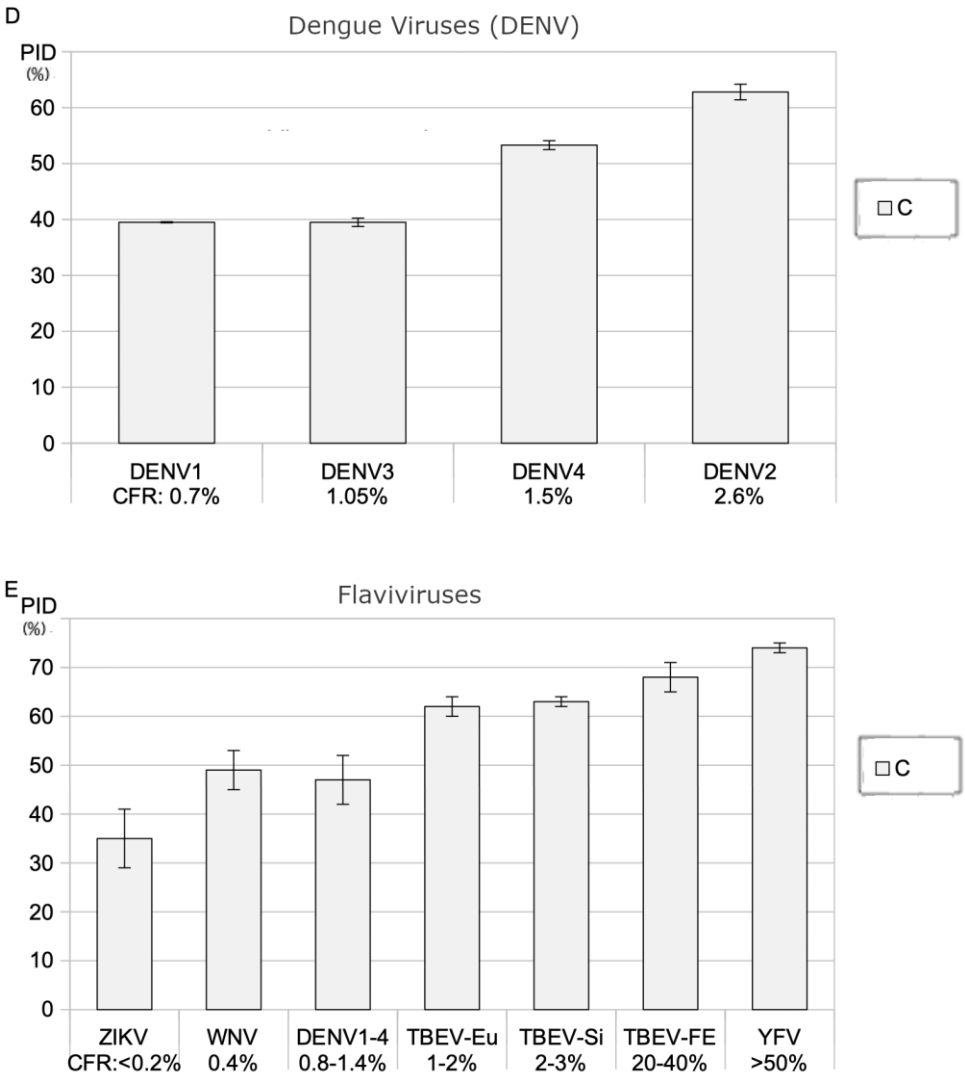
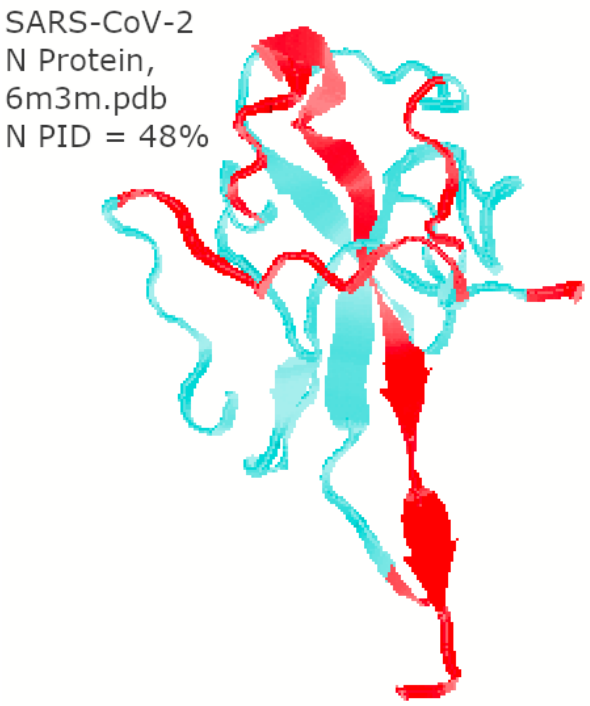


Figure 3



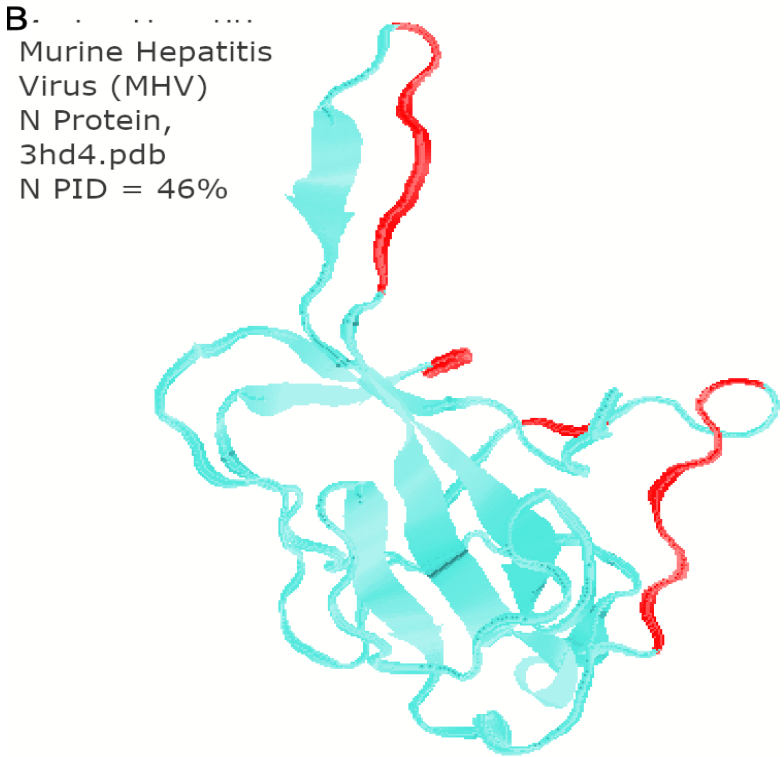
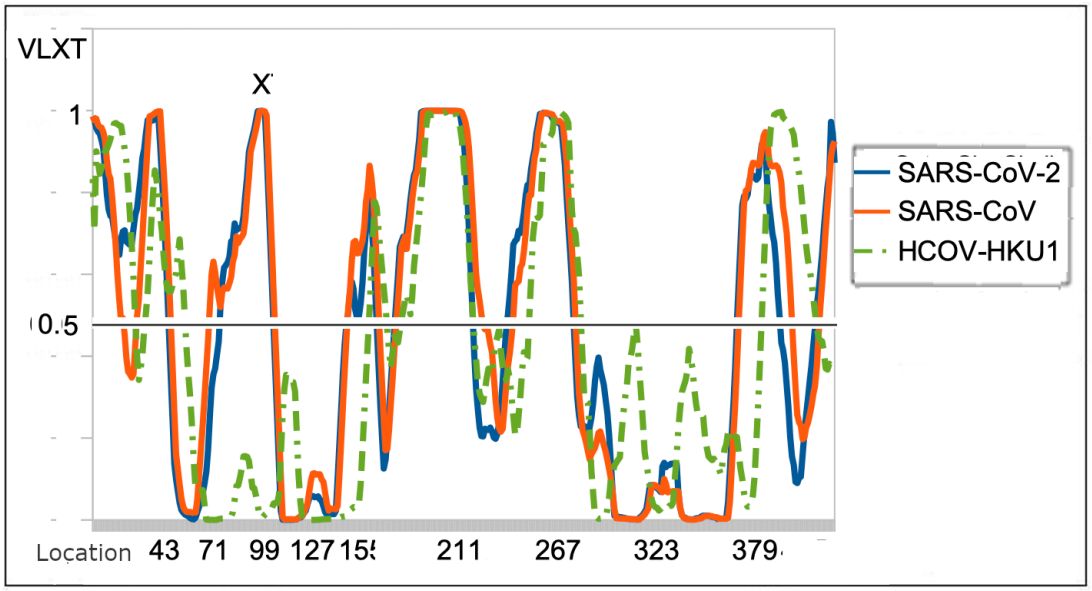


Figure 4

A



B

```
Location        1                                                          59
Disorder(D)     DDDDDDD DDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD
SARS-CoV-2      MSDNGPQ-NQRNAPRITFGGPSDSTGSNQNGERSGARSKQRRPQGLPNNTASWFTALTQH
                MSDNGPQ NQR+APRITFGGP+DST +NQNG R+GAR KQRRPQGLPNNTASWFTALTQH
SARS-C0V        MSDNGPQSNQRSAPRITFGGPTDSTDNNQNGGRNGARPKQRRPQGLPNNTASWFTALTQH
Disorder(D)     DDDDDDDDDDDDDDDDD           DDDDDDDDDDDDDDDDDDD
Location        1                                                          60
```

C

```
Location        41                                                         98
Disorder(D)     DDDD                                DDDDDDDDDDDDDDDDDDDDDDDDDDD
SARS-CoV-2      RPQGLPNNTASWFTALTQHGK-EDLKFPRGQGVPINTNSSPDDQIGYYRRATRR-IRGGD
                +PQG P    SWF+ +TQ  K  D KFP GQGVPI    P +  GY+ +  RR  +  D
HCOV-HKU1       QPQGNPIPHYSWFSGITQFQKGRDFKFPDGQGVPIAYGIPPSEAKGYWYKHNRRSFKTAD
Disorder(D)         NO DISORDERED RESIDUES  AT THIS LOCATION RANGE
Location        53                                                         112
```

Figure 5